Parveen Khurana

125 Followers About Following

Vanishing and Exploding Gradients

Parveen Khurana Jul 16, 2019 · 11 min read

This article covers the content discussed in the Vanishing and Expoding Gradients module of the Deep Learning course offered on the website: https://padhai.onefourthlabs.in

Taking a closer look at the derivative wrt W:

As discussed in the previous article, we know that the derivative of the loss function(at time step t) wrt W can be written as below:

$$\frac{\partial \mathscr{L}_t(\theta)}{\partial W} = \frac{\partial \mathscr{L}_t(\theta)}{\partial s_t} \sum_{k=1}^{t} \frac{\partial s_t}{\partial s_k} \frac{\partial s_k}{\partial W}$$

where

$$\frac{\partial s_t}{\partial s_k} = \frac{\partial s_t}{\partial s_{t-1}} \frac{\partial s_{t-1}}{\partial s_{t-2}} \frac{\partial s_{t-2}}{\partial s_{t-3}} \cdots \frac{\partial s_{k+1}}{\partial s_k}$$

that we have multiple paths from L4 to W so we need to sum the derivatives along all the possible paths from L4 to W.

The highlighted portion in the below image shows the derivative of loss(at time step t) wrt the hidden state(s) at the same time step ‘t’.

$$\frac{\partial \mathscr{L}_t(\theta)}{\partial W} = \frac{\partial \mathscr{L}_t(\theta)}{\partial s_t} \sum_{k=1}^{t} \frac{\partial s_t}{\partial s_k} \frac{\partial s_k}{\partial W}$$

And the part in the blue box sums the derivative along all the possible paths from the hidden state(at time step t) and we would have ‘t’ such paths so the summation goes from k = 1 to k = t. And at every time step, we have this derivative of st wrt sk and then the derivative of sk wrt W.

$$\frac{\partial \mathscr{L}_t(\theta)}{\partial W} = \boxed{\frac{\partial \mathscr{L}_t(\theta)}{\partial s_t}} \boxed{\sum_{k=1}^{t} \frac{\partial s_t}{\partial s_k} \frac{\partial s_k}{\partial W}}$$

Now taking a closer look at the derivative of st wrt sk: we can write this derivative as the following

$$\frac{\partial s_t}{\partial s_k} = \frac{\partial s_t}{\partial s_{t-1}} \frac{\partial s_{t-1}}{\partial s_{t-2}} \frac{\partial s_{t-2}}{\partial s_{t-3}} \cdots \frac{\partial s_{k+1}}{\partial s_k}$$

We can write that as a chain rule as s1 leads to s2, s2 leads to s3, s3 leads to s4, so we can break the same chain rule in the backward direction, so we would have this derivative of s4 wrt s1 as the following:

$$\frac{\partial \bar{s}_4}{\partial s_3} \frac{\partial s_3}{\partial s_2} \frac{\partial s_2}{\partial s_1}$$

And the number of elements in this chain would actually be the difference in the index of 4 and 1 i.e index/value of t and k.

So, the term in the image below

$$\frac{\partial s_t}{\partial s_k}$$

is actually a product of (t — k) terms

$$\frac{\partial s_t}{\partial s_k} = \frac{\partial s_t}{\partial s_{t-1}} \frac{\partial s_{t-1}}{\partial s_{t-2}} \frac{\partial s_{t-2}}{\partial s_{t-3}} \cdots \frac{\partial s_{k+1}}{\partial s_k}$$

And this product can be written compactly as

$$\frac{\partial s_t}{\partial s_k} = \prod^{t} \frac{\partial s_j}{\partial s_{j-1}}$$

So, if we have t = 20 and k = 1, then we would have the product of 19 such terms.

The formula for the derivative of st wrt sk is given as(in compact form):

$$\frac{\partial s_t}{\partial s_k} = \prod_{j=k+1}^{t} \frac{\partial s_j}{\partial s_{j-1}}$$

Let’s take a look at one term from the above formula:

$$\frac{\partial s_j}{\partial s_{j-1}}$$

Fig 1

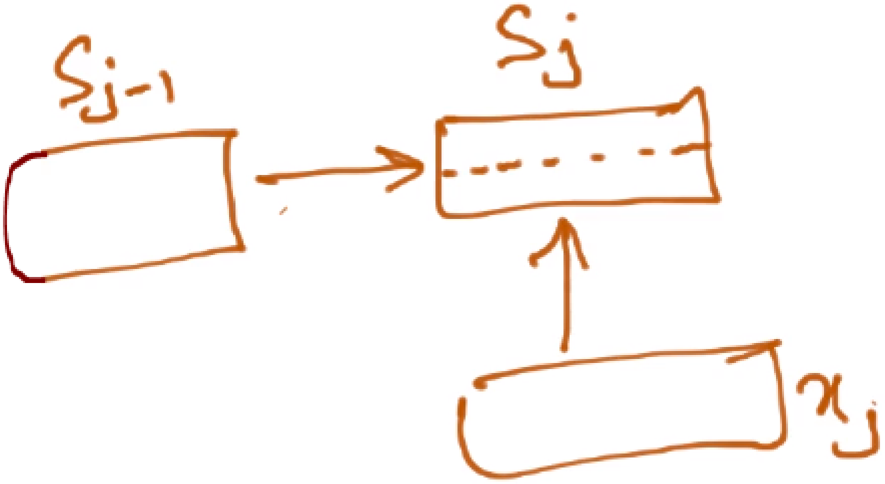
And we also have the equations for pre-activation and the activation as the following:

$$a_j = Ux_j + Ws_{j-1} + b$$
$$s_j = \sigma(a_j)$$

Fig 2

Now based on the relations in Fig 1 and Fig 2, we can write the value of the derivative depicted in Fig 1 as the following:

$$\frac{\partial s_j}{\partial s_{j-1}} = \frac{\partial s_j}{\partial a_j} \frac{\partial a_j}{\partial s_{j-1}}$$



aj and sj would be represented as:

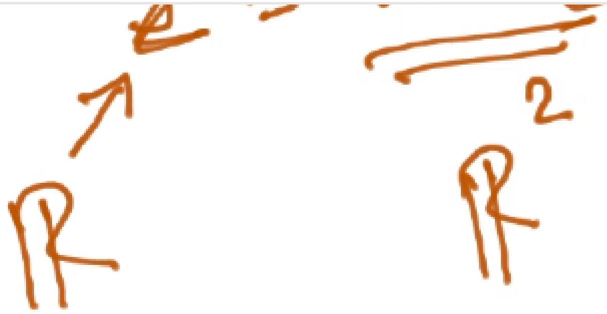
$$\begin{aligned} a_j &= [a_{j1}, a_{j2}, a_{j3}, \ldots a_{jd},] \\ s_j &= [\sigma(a_{j1}), \sigma(a_{j2}), \ldots \sigma(a_{jd})] \end{aligned}$$

Every element of sj depends on the corresponding element of aj, it does not depend upon all the elements of aj(as we have used element-wise sigmoid function), for example, sj1 depends only on aj1

$$s_{j1} = \frac{1}{1+e^{-a_{j1}}}$$

Let’s take one example:

Let z be equal to x square plus y square



Now we collectively store x and y in theta, so z becomes a function of theta

$$\theta = [x, y]$$
$$z = f(\theta)$$
$$= f(x, y)$$

Now if we take the derivative of z wrt theta, that would give us two terms(so we say that this derivative is going to be 2 dimensional as theta is 2 dimensional):

$$\left[\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y} \right]$$

So, the quantity whose derivative we are computing(z in this case) is a scalar and the quantity wrt which we are computing the derivative(theta in this case) is a vector, then the derivative is going to be a vector.

$$z_1 = x^2 + y_2$$

$$z_2 = x^3 + y_3$$

$$z = [z_1 \quad z_2]$$

$$\theta = [x, y]$$

If we compute the derivative of z wrt theta that would be:

$$\begin{bmatrix} \frac{\partial z_1}{\partial x} & \frac{\partial z_2}{\partial x} \\ \frac{\partial z_1}{\partial y} & \frac{\partial z_2}{\partial y} \end{bmatrix}^T$$

In short, if we take the derivative of 2-dimensional quantity(z) wrt a 2-dimensional quantity(theta), then the derivative would be the derivative of every element of numerator wrt to every element of the denominator, so in this, our derivative would be of size: 2 X 2.

Having done this, we will use this concept in our problem of computing derivative of sj wrt aj

$$\frac{\partial s_j}{\partial s_{j-1}} = \boxed{\frac{\partial s_j}{\partial a_j}} \frac{\partial a_j}{\partial s_{j-1}}$$

Now, in this case, sj is a ‘d’ dimension vector and ‘aj’ is also a ‘d’ dimension vector

$$a_j = [a_{j1}, a_{j2}, a_{j3}, \ldots a_{jd},]$$
$$s_j = [\sigma(a_{j1}), \sigma(a_{j2}), \ldots \sigma(a_{jd})]$$

And based on the above concept, we can say that the derivative of sj wrt aj is going to be a ‘d X d’ matrix. And the elements of this matrix would look like(it would be derivative of every element of sj wrt every element of aj):

$$\frac{\partial s_j}{\partial a_j} = \begin{bmatrix} \frac{\partial s_{j1}}{\partial a_{j1}} & \frac{\partial s_{j2}}{\partial a_{j1}} & \frac{\partial s_{j3}}{\partial a_{j1}} & \cdots \\ \frac{\partial s_{j1}}{\partial a_{j2}} & \frac{\partial s_{j2}}{\partial a_{j2}} & \ddots & \\ \vdots & \vdots & \vdots & \frac{\partial s_{jd}}{\partial a_{jd}} \end{bmatrix}$$

Now let’s take sj2, its formula would be:

So, if we take the derivative of sj2 with respect to anything except aj2, then the derivative is going to be 0.

So, taking the above point into consideration, the derivative would look like

$$\frac{\partial s_j}{\partial a_j} = \begin{bmatrix} \frac{\partial s_{j1}}{\partial a_{j1}} & \frac{\partial s_{j2}}{\partial a_{j1}} & \frac{\partial s_{j3}}{\partial a_{j1}} & \cdots \\ \frac{\partial s_{j1}}{\partial a_{j2}} & \frac{\partial s_{j2}}{\partial a_{j2}} & \ddots & \\ \vdots & \vdots & \vdots & \frac{\partial s_{jd}}{\partial a_{jd}} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma'(a_{j1}) & 0 & 0 & 0 \\ 0 & \sigma'(a_{j2}) & 0 & 0 \\ 0 & 0 & \ddots & \\ 0 & 0 & \cdots & \sigma'(a_{jd}) \end{bmatrix}$$

This a diagonal matrix with all off-diagonal elements as 0

So, we can write it like the below:

$$\frac{\partial s_j}{\partial a_j} = \begin{bmatrix} \frac{\partial s_{j1}}{\partial a_{j1}} & \frac{\partial s_{j2}}{\partial a_{j1}} & \frac{\partial s_{j3}}{\partial a_{j1}} & \cdots \\ \frac{\partial s_{j1}}{\partial a_{j2}} & \frac{\partial s_{j2}}{\partial a_{j2}} & \ddots & \\ \vdots & \vdots & \vdots & \frac{\partial s_{jd}}{\partial a_{jd}} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma'(a_{j1}) & 0 & 0 & 0 \\ 0 & \sigma'(a_{j2}) & 0 & 0 \\ 0 & 0 & \ddots & \end{bmatrix}$$

$$= diag(\sigma'(a_j))$$

which conveys that it is a diagonal matrix whose all the elements are this sigma-dash aj where this sigma-dash aj is a collection of the terms sigma-dash aj1, sigma-dash aj2 and so on(and this sigma-dash represents the derivative of the sigmoid).

$$\frac{\partial s_j}{\partial s_{j-1}} = \boxed{\frac{\partial s_j}{\partial a_j}} \frac{\partial a_j}{\partial s_{j-1}}$$

So, the first part(red box in the above image) is computed and it is equal to a diagonal matrix as derived above. Now let's look at the second part

The relation between aj and s(j-1) is given as

$$a_j = Ux_j + Ws_{j-1} + b$$

The first term in the above relation is xj which has no dependency on s(j-1), b does not have any dependency on s(j-1)

So, the derivative of aj wrt s(j-1) would be W.

You can also look it like: aj is a 'd' dimension vector, s(j-1) is a 'd' dimension vector so the derivative of aj wrt s(j-1) is going to be a 'd X d' matrix and W indeed is a 'd X d' matrix

$$\frac{\partial s_j}{\partial s_{j-1}} = diag(\sigma'(a_j))W$$

the update would be very large and if the above term is very small then the update would be a small quantity. Both the cases are not desirable as if it is a small quantity then we are not moving much from the original value of W and if it a large quantity then let’s say we are at a particular weight/value and we are moving suddenly by a very large quantity.

$$w = w - \eta \frac{\partial L}{\partial w}$$

So, we want to compute the magnitude of this derivative of sj wrt s(j-1) as the parameter updation depends on this

magnitude of $\frac{\partial s_j}{\partial s_{j-1}}$

$$\left\|\frac{\partial s_j}{\partial s_{j-1}}\right\| = \|diag(\sigma'(a_j))W\|$$

$$\leq \|diag(\sigma'(a_j))\|\|W\|$$

$$\left\|\frac{\partial s_j}{\partial s_{j-1}}\right\| = \|diag(\sigma'(a_j))W\|$$

$$\leq \|diag(\sigma'(a_j))\|\|W\|$$

Let’s compute the magnitude of the underlined part in the above image, now aj is a ‘d’ dimension vector and we are taking the element-wise derivative of it, the magnitude of the matrix would depend upon the magnitude of the individual elements in it. And the individual elements are the derivative of the sigmoid(or tanh) function.

Now the derivative of this sigmoid of aj can be written as:

$$\sigma'(a_j) = a_j(1-a_j) = a_j - a_j^2$$

Now the maximum value of the quantity in the above image can be computed by taking the derivative and equating it to 0 which gives the value of aj for which the above quantity(derivative of sigmoid of aj) would be maximum.

So, if we take derivate we get (1–2*aj = 0) which gives aj as (1/2) and if we put this value of aj in the main quantity we get the output as (1/2–1/4) = (1/4) and this is in the case of the logistic function.

In the case of tanh non-linearity, we have

$$\sigma'(a_j) = 1-(a_j)^2$$

So, again we take the derivative of the above quantity to find the value of aj for which the above quantity would be maximum. So, the derivative of above quantity wrt aj

$$\begin{aligned}\sigma'(a_j) &\le \frac{1}{4} = \gamma\ [if\ \sigma\ is\ logistic]\\ &\le 1 = \gamma\ [if\ \sigma\ is\ tanh]\end{aligned}$$

Since the elements of the matrix are bounded that implies the magnitude of diagonal itself would be bounded.

Now the equation below

$$\begin{aligned}\left|\left|\frac{\partial s_j}{\partial s_{j-1}}\right|\right| &= ||diag(\sigma'(a_j))W||\\ &\le ||diag(\sigma'(a_j))||\,||W||\end{aligned}$$

could be re-written as

$$\left|\left|\frac{\partial s_j}{\partial s_{j-1}}\right|\right| \le \gamma||W|| \le \gamma\lambda$$

As W is the weight matrix that we initialize initially(we won’t initialize it to infinite value), and would be in some limit(bounded) and would remain in some limit even after the updates, let represent this range/limit by lambda so the overall equation can be written as:

$$\left\|\frac{}{\partial s_{j-1}}\right\| \leq \gamma\|W\| \leq \gamma\lambda$$

Our original derivative has the derivative of st wrt to sk and in that, we came up to the part of derivative of sj wrt s(j-1)

So, our original derivative is

$$\left\|\frac{\partial s_t}{\partial s_k}\right\| = \left\|\prod_{j=k+1}^{t} \frac{\partial s_j}{\partial s_{j-1}}\right\|$$

And this derivative of st wrt sk is the summation of many such derivatives(right-hand side of the above equation) and for each of those derivatives, some limit is there and we know that their maximum value could be (gamma * lambda)

$$\begin{aligned}\left\|\frac{\partial s_t}{\partial s_k}\right\| &= \left\|\prod_{j=k+1}^{t} \frac{\partial s_j}{\partial s_{j-1}}\right\| \\ &\leq \prod_{j=k+1}^{t} \gamma\lambda \\ &\leq (\gamma\lambda)^{t-k}\end{aligned}$$

So, the overall value of the derivative of st wrt sk would be less than equal to the value in the red box(in the below image)

$$
\begin{aligned}
\left\|\frac{\partial s_t}{\partial s_k}\right\| &= \left\|\prod_{j=k+1}^{t}\frac{\partial s_j}{\partial s_{j-1}}\right\| \\
&\leq \prod_{j=k+1}^{t}\gamma\lambda \\
&\leq \boxed{(\gamma\lambda)^{t-k}}
\end{aligned}
$$

So the overall value of derivative of st wrt sk is going to be a product of many terms and all of these are in some limit, so if these terms are small then the product is going to be even smaller than that and if these terms are in the normal range even then the product of so many terms would make the overall value very large. For example. if each of these value is let's say is 1/2 and 10 such terms are there that would make the final value as (0.5 raised to power 10) and in the second case if we assume the value of each of these term as 2 then the final value would be (2 raised to power 10 = 1024).

So, the overall point is:

While computing the value of derivative of loss function wrt W, we have one term corresponding to the derivative of st wrt sk, on analyzing further, we got to know that this derivative of st wrt sk is actually a product of many such terms so we analyzed one such term and we arrived at the conclusion that each such term lies in a range and is bounded and its maximum value could be (gamma * lambda), so if we have many such terms then our upper limit(maximum value) would be (gamma * lambda) raise to power (t — k).

So, if we have (gamma * lambda) in the range 0 to 1, then a very high power of that(gamma * lambda) would diminish the final value to 0 and this would result in the problem of vanishing gradient; on the other hand, if the value of (gamma * lambda) is greater than 1 then the overall quantity would be very large and this would result in exploding gradients. So, this is one of problem associated with training RNNs.

All the images used in this article is taken from the content covered in the Vanishing and Exploding module of the Deep Learning Course on the site: padhai.onefourthlabs.in

Open in app

About Write Help Legal